Automating work in Galaxy

Tom Doak
Le-Shin Wu
Carrie Ganote
National Center for Genome Analysis Support

August 12, 2015

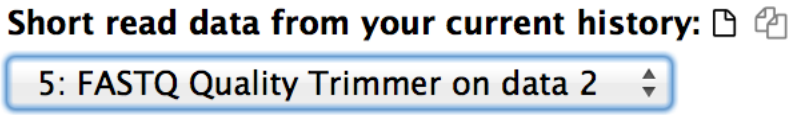
INDIANA UNIVERSITY

Automation Overview

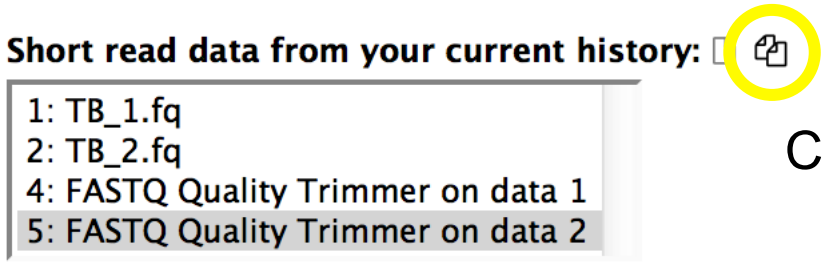
There are two main approaches to automating work in Galaxy:

- Running one tool on lots of inputs
- Creating and running workflows

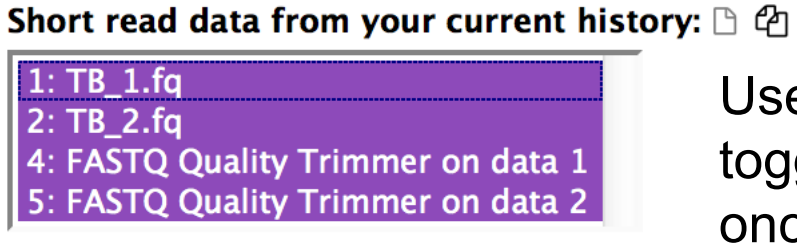
Multiple inputs in one tool



Some tools, such as FastQC, support multiple inputs. Note that these will run as separate, unrelated jobs!

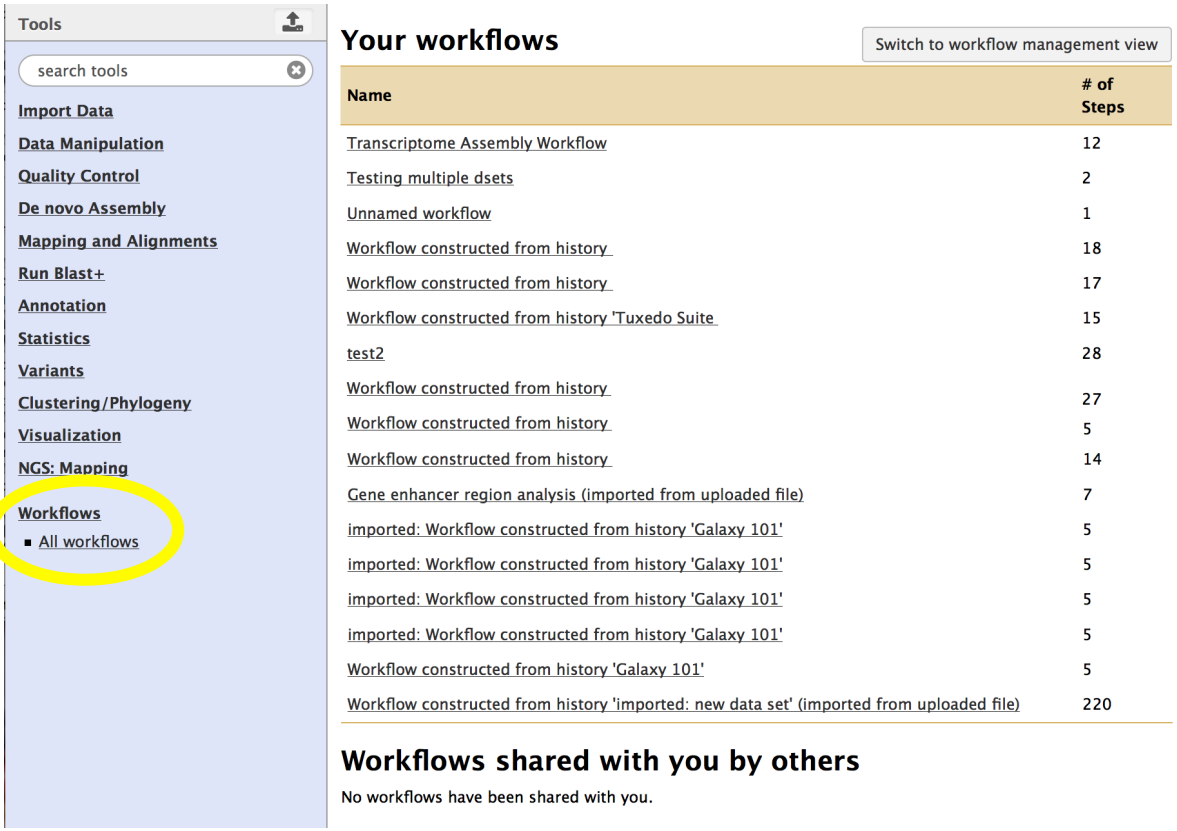


Click on “Run tool in parallel..”



Use the shift and control keys to toggle or select many inputs at once.

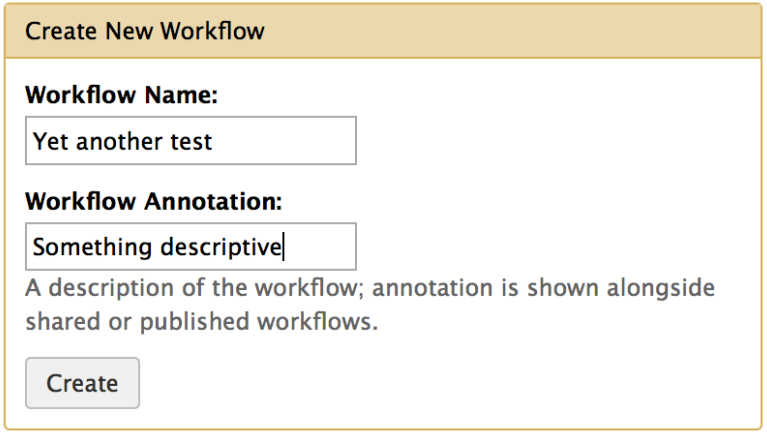
Introduction to workflows



Workflows encapsulate an entire histories worth of work. They can be created by:

- Building from scratch
- Extracting from history
- Importing from shared workflow

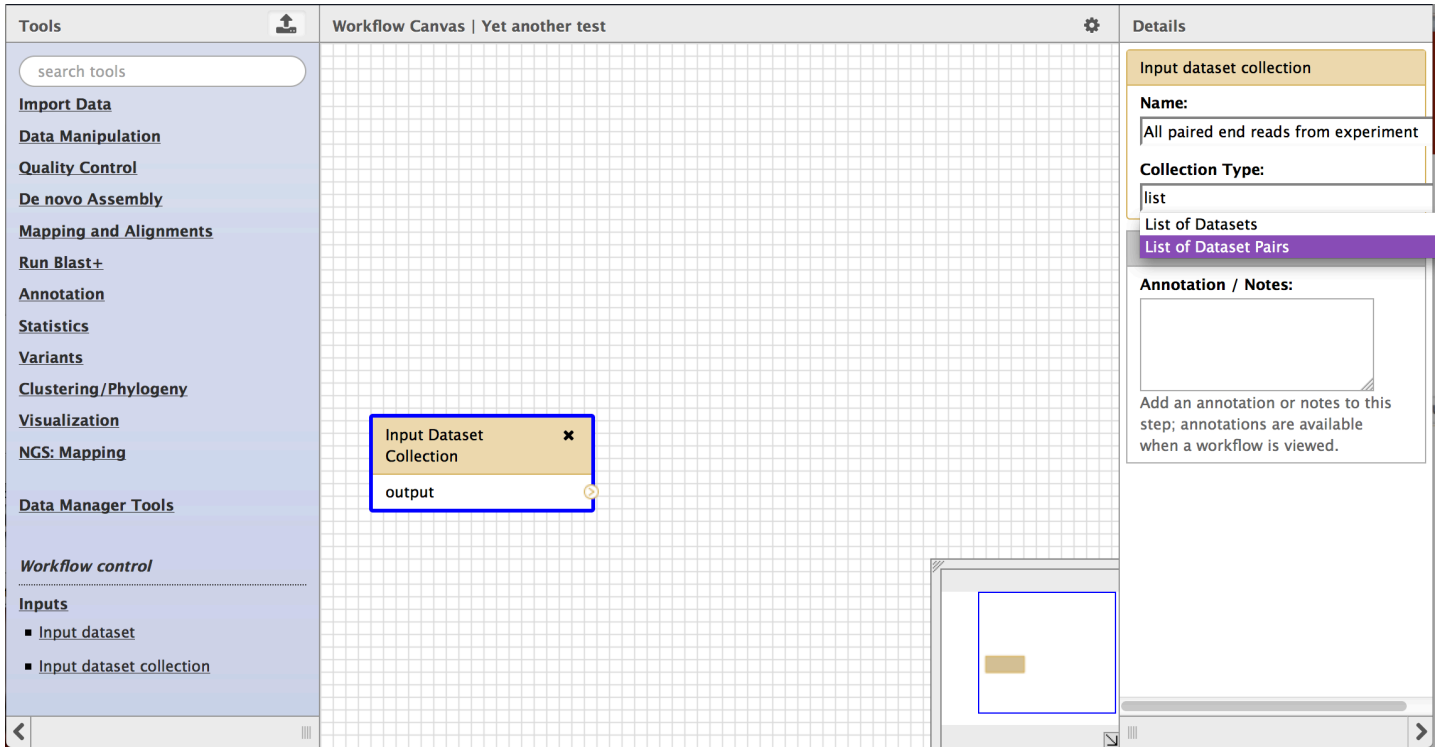
Building workflows



Create a new workflow and annotate it with something you will remember a year from now!

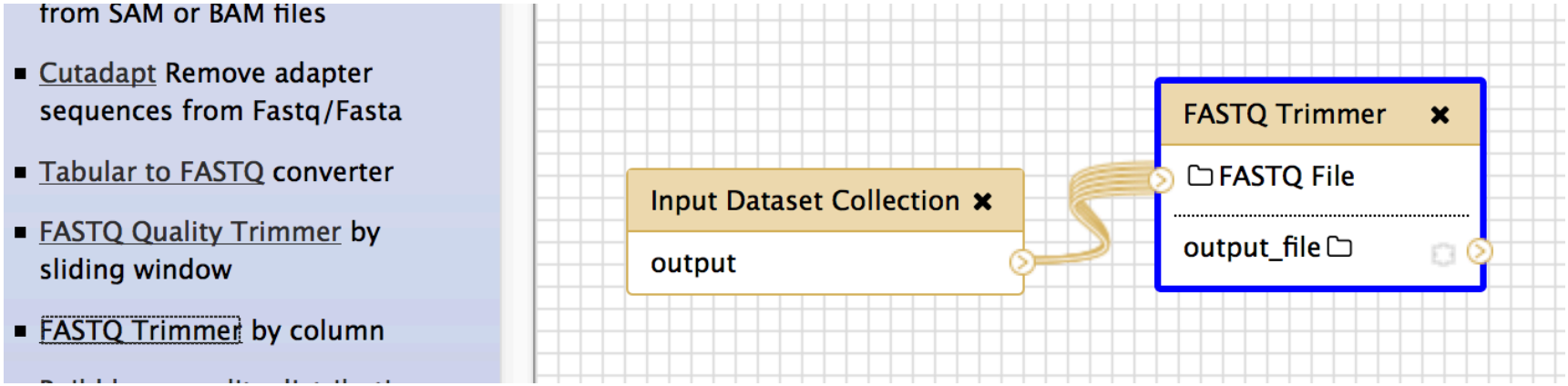
Building workflows

Each step is created by choosing tools to build the pipeline. Here is an example input with a dataset collection.

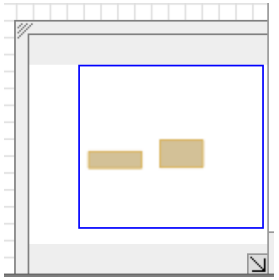


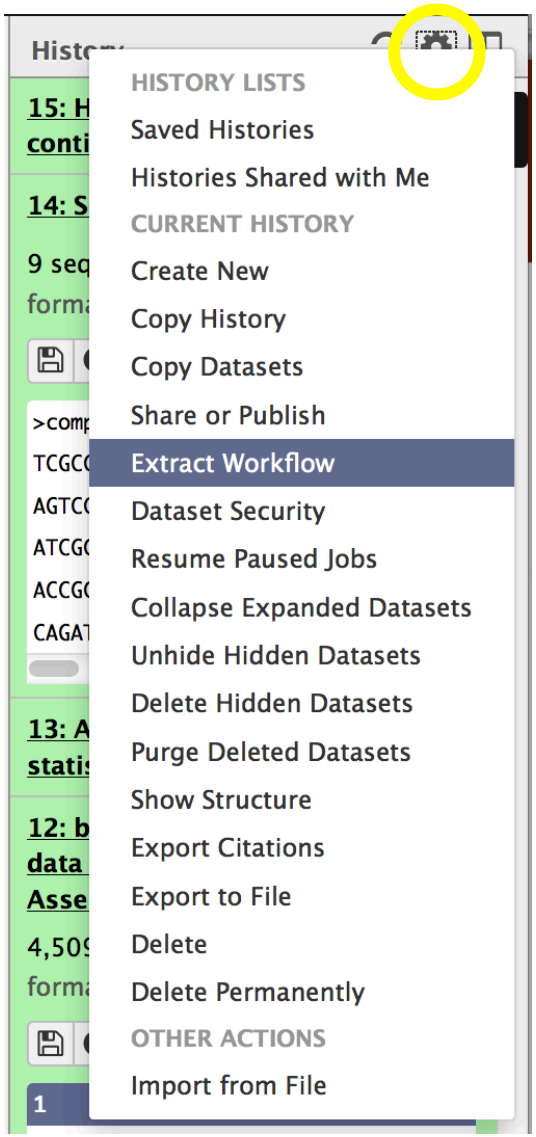
Building workflows

Outputs from one tool can be clicked and dragged into the input of another tool. This process can be extended until the entire pipeline is created.



The bottom right side has an overview panel to allow the entire board to be scrolled by clicking and dragging.





Extracting workflows

An entire history can be converted into a workflow for easy sharing and to reproduce work. Make sure you are in the history that you want to extract from.

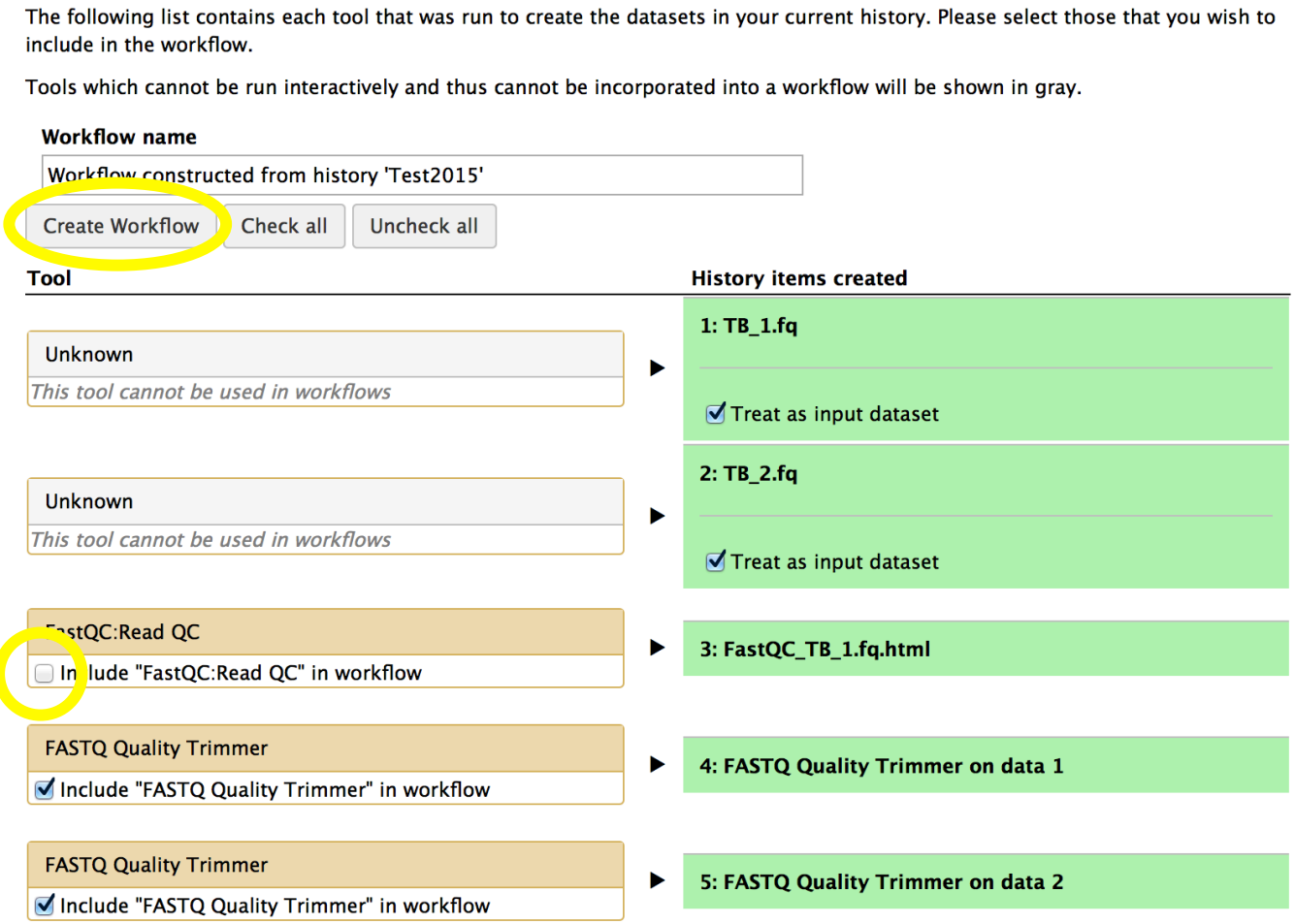
Extracting workflows

As always, a descriptive name will help in the long run!

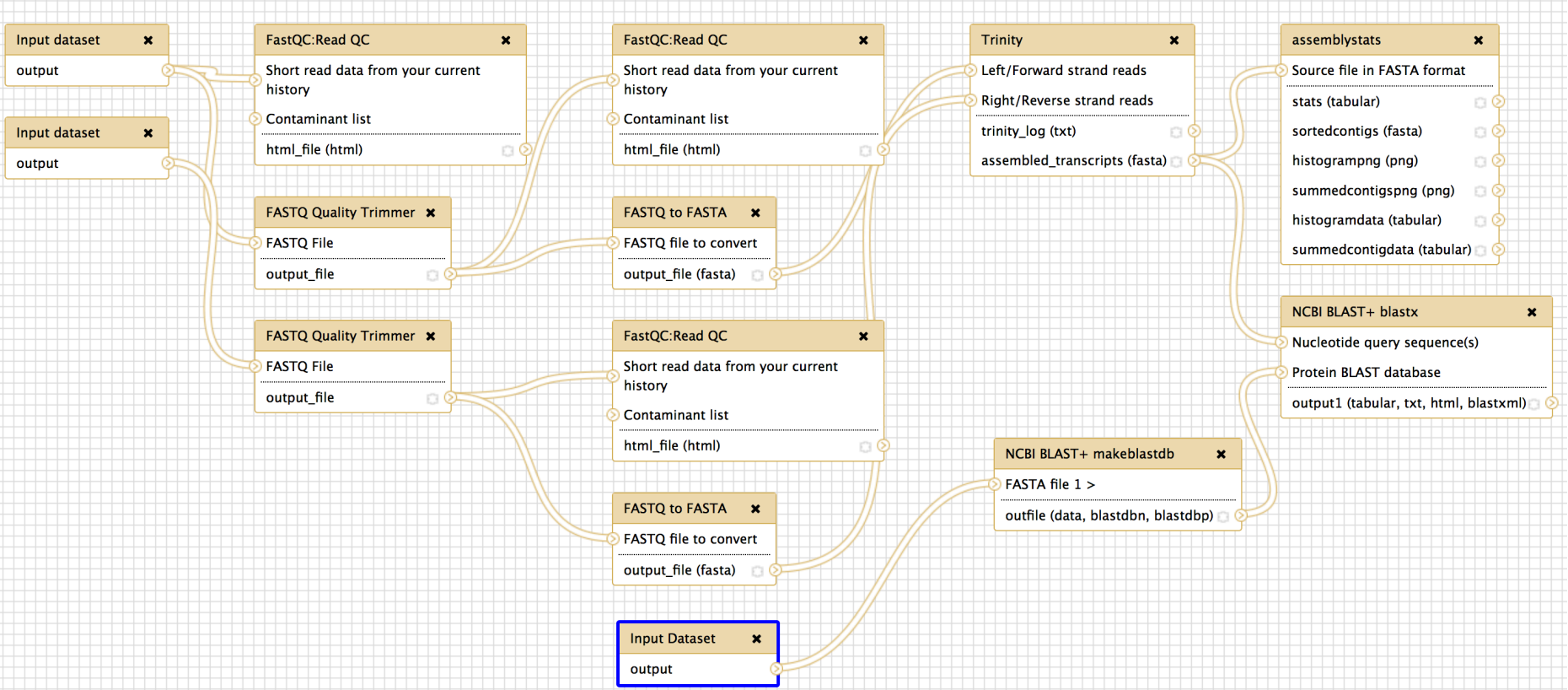
Create Workflow when done

Input files are chosen later.

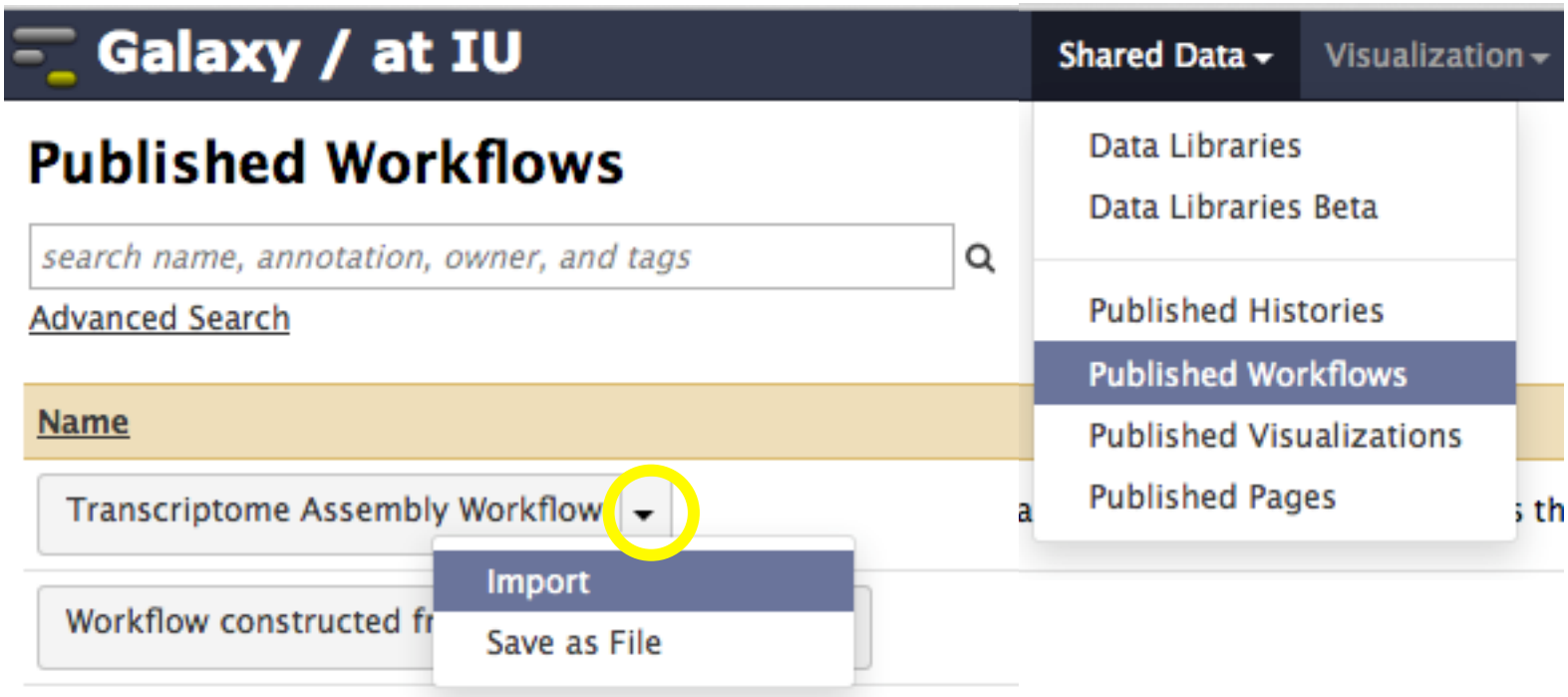
You can skip steps you don't need



Extracting workflows



Importing workflows

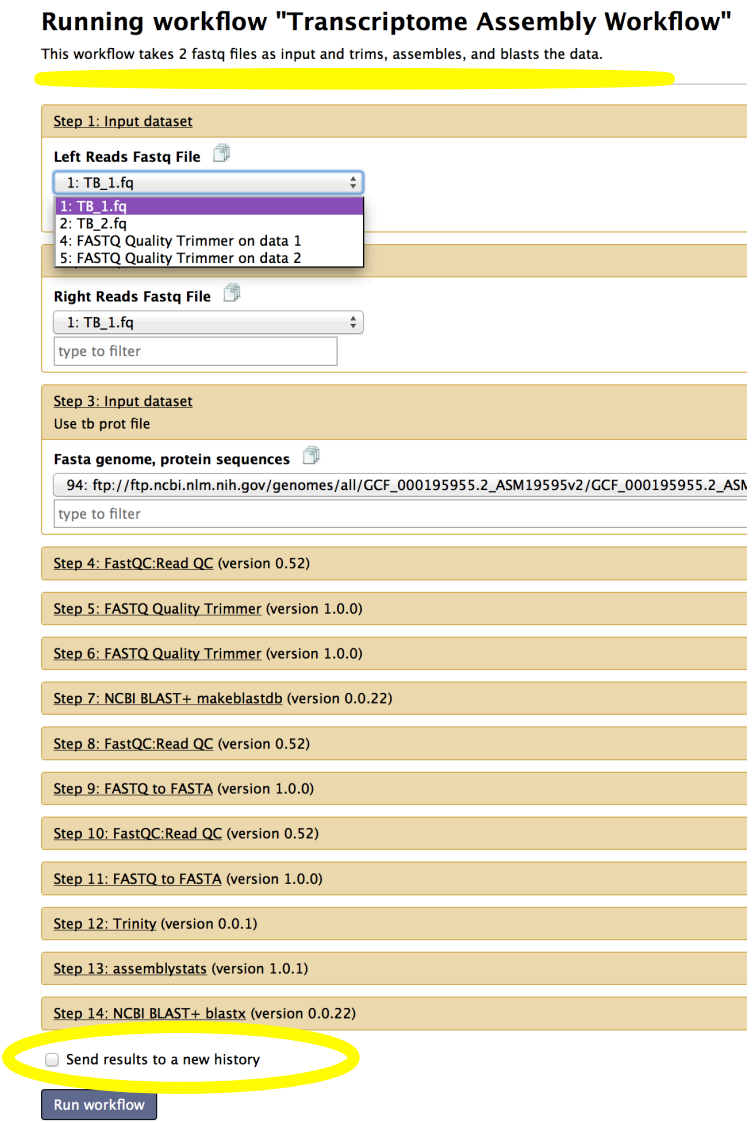


Sharing workflows makes it easy for others to reproduce your work and for you to run new inputs through the same procedure

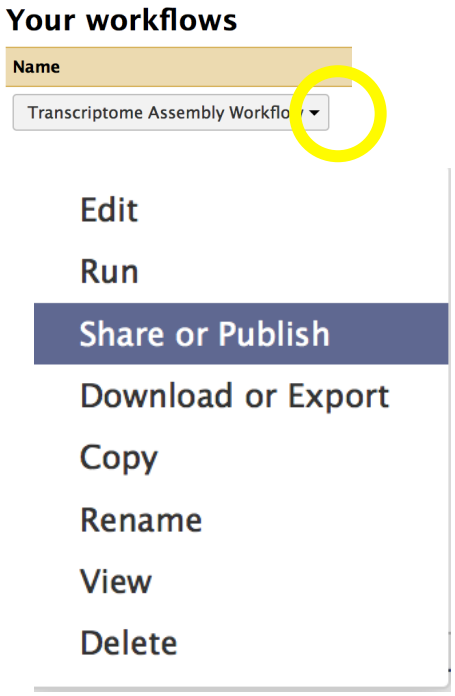
Running workflows

The inputs are the only thing you need to set up in order to run the workflow.

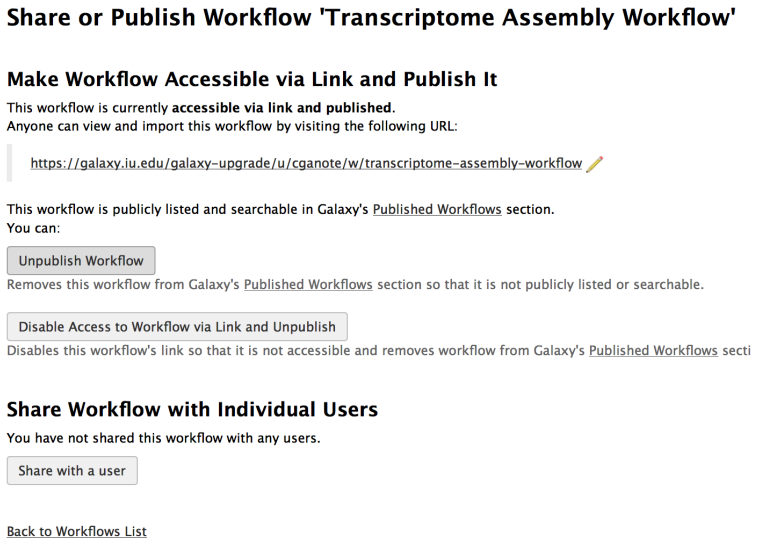
For a cleaner history, send results to a new history:



Sharing workflows



Sharing workflows makes it easy for others to reproduce your work and for you to run new inputs through the same procedure



Future directions

There are other methods in development that will add even more flexibility to running massive numbers of jobs and increasing throughput:

- API calls for advanced use
- Dataset collections
- Nested workflows

Fín

Thanks for watching!

Questions and comments:

Email help@ncgas.org